(U) *Ask Raul*: Dictionary Equations

FROM: 'Raul', a DNI Analyst
Unknown
Run Date: 09/08/2004

Dear Raul,

(S) Why is it that my dictionary category hauls back such a huge mess? I've tried everything I can think of and all that I have is a disaster. What gives?

-- Bob

Bob,

(S) You know, there is probably no better place than in the DNI world where the old saying "Physician, heal thyself" is more true. Old Raul has been around the block a few times and the vast majority of the pain and suffering in the DNI world is self-inflicted.

(S) Let's look at a sample of one of the ingenious ways we analysts go about tasking things in the dictionaries. Here it is:

('AEROSOL' AND 'GENERAT')

Now, looking at this and knowing which category it came from, the analyst was attempting to get things like: aerosol generator ... aerosol generation ... generation of aerosols, etc.

Unfortunately, the analyst forgot about or didn't consider a few key issues, such as:

1. (S) Dictionaries look for the search terms as simple patterns. So, 'GENERAT' would hit on:

   GENERATE
   GENERATES
   GENERATED
   GENERATING
   GENERATOR
   GENERATION
   GENERATE
   GENERATES
   GENERATOR
   REGENERATE
   REGENERATES
   REGENERATED
   REGENERATING
   REGENERATOR
   REGENERATION
   REGENERATE
   REGENERATES
   REGENERATOR, etc.

   Did the analyst take this into account? Apparently not. This happens all the time and creates a lot of problems. An easy

SERIES:
(U) Ask Raul - Answers to DNI Questions

way out, in this case, would have been to include whitespace, such as:

' GENERAT'

'\0x0aGENERAT'

It doesn't solve the whole problem but it knocks a lot of stuff out instantly.

2. (S) Logical statements have to be logical. Seems rather apparent, but quite often what an analyst puts into the dictionaries makes no real sense, logically. Using this example again, the 'AND' means this equation will produce hits if the pattern 'AEROSOL' appears anywhere and the pattern 'GENERAT' appears anywhere. So, I could have this in traffic: *"aerosol generation unit for use by terrorists to kill Americans"* ... or ... *"aerosol spray hair conditioner"* ... (many, many Kbytes before or after) ... the song, *"Talkin' 'Bout My Generation"* ...

   This equation would result in both of these items being returned by the category. Likewise, using an equation like:

   'AEROSOL' ***WITHIN*** 'GENERAT'

   ...is hardly any better. If what I wanted was to get anything which contains 'AEROSOL' followed by the pattern 'GENERAT' anywhere after 'AEROSOL' then this is the way to go. Generally speaking, this is not what most folks want. They want, for example, 'AEROSOL' fairly close to 'GENERAT' not 'AEROSOL' at the beginning of the session and 'GENERAT' thousands or millions of bytes away.

   However, using a **WITHIN** with a number would be much better. For example:

   'AEROSOL' ***WITHIN 12*** 'GENERAT'

   This allows for a good bit of slop between the two terms but at least puts them in close proximity to each other. If combined with the above suggestion from item 1, things will be even better. If I do this, I have a reasonable chance of getting the sessions actually mentioning *"aerosol generators"* and such, as opposed to sessions about anything under the sun containing those two patterns like a web page made by "Microsoft HTML GENERATor" with an ad for an "AEROSOL hairspray".

3. (S) In virtually all cases, the scan_as tables for the dictionaries are set up to be case insensitive. So, things like the following would produce valid hits on the term **'GENERAT'** :

   GENERATE
   generate
   Generate
   GeNeRaTe
   GENerAte
   genERaTe
   etc.

   Can you find the term 'GENERAT' in the lines below? Recognize them as base64?

XrYGKI4MIGr/bvBEkBppxPQWwnZKgeNerAT6ucgCFYWEAEZ
bcHIEBB1Tej0ktCDTODHcJZ6ditwBTsKziwxMDCDY0jD7O/B0
jYCjXmp1uzN4fRCzT3CD6DEQwoTJ+BJ0Lg4kGENerATS0fs5
OHNDDvIdOneQegenERATzrw72LwRMICtlgSBVS6TPUhEYOO
cjU1I8IGXps1GMPhFdGqzlouLkQXfkm39YE65L8EjHrdRxLdkZ
9AfttAPc2hsqaW6hkzgnn5ABDB7MxuYUoGENeRaTlw0iW86x
tmS4dLJHJA3nQkBPgQclbwjke3fZjl/0IoBiEQV4PDQgy+Pgen
eRATudCR8NTrktDGRAVSsiAnW97A0Ct9a3Dbjr48u

(S)This too has to be considered especially if your traffic is being collected by a system which cannot handle encoding and which does not process it prior to the selection and filtering process.

4. (S) In many, many cases the analyst throws in a very generic term then, when it becomes obvious there is too much junk, begins defeating things instead of positively selecting on what he wants. This little area is quite interesting. A typical example old Raul has seen over and over goes like this -- an analyst puts a selector into a category such as:

   'ABDULBADGUY@HOTMAIL'

   (S) Well, in short order the analyst is crying over his traffic because it is basically a mountain of spam and other worthless stuff. So, does the analyst do some analysis and then try to include terms to get what he wants? Absolutely not! He looks at a handful of items and notices there are several press items. So, he sticks 'PRESS' into a defeat and thinks everything is swell. The problem is, he doesn't realize 'PRESS' will defeat: IMPRESS, REPRESS, SUPPRESS or "PRESS THE BUTTON TO ACTIVATE THE BOMB", etc.

   (S) Next thing you know the analyst has a defeat section which is larger than everything else in the category, and a few minutes of review -- much to the surprise and horror of the analyst -- shows that those defeat terms are defeating him right out of the very traffic he wants.

   (S) So, instead of going ape with defeats, look at your traffic and go after what you want. If you don't know what ICMP is, don't defeat it out of your category, use positive selectors to go after the traffic type you want. Only interested in ports 25 and 110? Well, include them in your equations to make sure that is what you get. Think positive!

(S) There you have it. Yes, the system is not perfect, I've skimmed over these topics and there is much, much more to be discussed but heck, given the way we analysts are doing things, even a perfect system wouldn't work. The next time you find yourself crying over the big DNI mess you have, look in the mirror first. You'll be looking at the person most responsible for it.

Raul